Evaluating a Model of Global Psychophysical Judgments for Brightness: I. Behavioral Properties Summations and Productions

Ragnar Steingrimsson
Department Cognitive Sciences
University of California, Irvine

Ragnar Steingrimsson
Department of Cognitive Sciences
University of California, Irvine
Irvine, CA, 92697-5100
e-mail: ragnar@uci.edu

Abstract

Steingrimsson and Luce (2005a,b, 2006, 2007) evaluated Luce's (2002, 2004) proposed psychophysical theory in loudness and found it substantially supported. The aim of the current research is to begin an extension of this research to brightness. Luce's theory deals with the global percept of subjective intensity, in which there is a psychophysical function Ψ that maps pairs of physical intensities onto the positive real numbers and represents, in an explicit mathematical way, subjective summation and a form of ratio production. These representations derive from a number of behavioral properties including certain plausible background assumptions. In three experiments involving the subjective perception of luminance, brightness, the key behavioral properties of summation over the two eyes and a form of generalized ratio production are empirically evaluated. Considerable support is reported for particular forms of Ψ for summations and ratio productions separately.

In a series of four papers, Steingrimsson and Luce (2005a,b, 2006, 2007) evaluated Luce's (2002, 2004) theory of global psychophysics in loudness. The results provided broad support for the branch of the theory in which the left and the right ears are not assumed to be behaviorally completely alike (the *biased* or *asymmetric case*). The main aim of this paper is to begin an analogous series of evaluations of the theory interpreted for the brightness domain. The paper is organized as follows:

• Relevant theory and interpretation in brightness: Summarizes the relevant portions of Luce's (2002, 2004) theory of global psychophysical judgments. In the tradition of the axiomatic approach, the theory presents non-domain-specific elements (primitives) which

are interpreted in the context of brightness. To place this paper in that context, the on-going experimental program is outlined.

• Experiments: Three experiments are presented, in which three behavioral properties, or axioms, are empirically evaluated. What about the theory can be concluded on the basis of the results is discussed.

• Summary, conclusion, and further work: The paper's subject matter and overall conclusions are summarized and the consequences of those conclusions for subsequent work are outlined.

• Appendices address the historical background of the work, the axiomatic psychophysical approach, and how Luce's (2002, 2004) model fits in with previous and related work. To accommodate a variety of readers, these are presented as separate appendices to be read as needed.

Summary of theory and interpretation in brightness

Brief overview of experimental setting and stimulus

Experiments are carried out using a computer monitor and a keyboard with which subjects, seated in dark room, can alter the luminance of the stimuli in accordance with instructions. The stimuli consist of squares of achromatic light—see the first panel in Figure 2 for an example. The tasks consist of making a certain stimulus equal in brightness to a standard (*matching*) or to some proportion of a standard, e.g., two times as bright as a standard (called either *ratio or magnitude production*). In some cases, a stereoscope—see middle panel in Fig. 2—is used to generate the stimuli.

Whatever physical signal attribute is taken to correspond to the sensation we call brightness, it seems crucial that it be defined independent of a hypothetical observer (otherwise, the physical stimulus is mutable by change in observer, a property not accorded variables such as temperature, volume, length, etc.). Here, brightness and its change are taken to be the sensations elicited by a luminance and changes of it. This functional definition of brightness accords with a great many other experiments (see Grossberg & Kelley, 1999, see also Ding & Sperling, 2006). Intuitively, brightness is the sensation that changes when, e.g., we vary the luminance setting on a computer monitor, or go from a sun-filled day into a movie theater.

It should be stressed that since Luce's (2002, 2004) theory is non-domain specific, it is applicable to any domain in which its primitives may be defined. Thus, the theory is not one of binocular brightness, but rather the theory is applied to binocular brightness. For that reason, when we refer to the primitives in the context of binocular brightness, it

This research was supported in part by National Science Foundation grants SBR-9808057 & BCS-0720288 (R. D. Luce, PI) to the University of California, Irvine—any opinion, finding, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation. Additional financial support was provided by the School of Social Sciences and the Department of Cognitive Sciences at UC Irvine. I am especially grateful to Marisa Carrasco for unfettered access to her laboratory and equipment at NYU to carry out much of the research reported here. I thank Randolph Blake and Jack Yellott for valuable input on technical matters. And I appreciate many helpful comments of Joetta Gobell and R. Duncan Luce on earlier versions as well as those of Michael Rudd who, in his role as a reviewer, made we work a lot but in return for a much improved paper.

is only due to the specific application of the theory here to that domain—further details in Appendix B.

Primitives

The first step in the axiomatic approach is the specification of the theory's primitives. Interpreting these in the context of brightness will determine the stimuli as well as how they may be manipulated (i.e., methods)—details in the first section of Appendix B

Joint presentations. The first primitive is the set of ordered pairs (x, u), where x and u correspond to physical intensities. Our interpretation of this primitive in the visual domain is that of squares of achromatic light (RGB guns set at equal value) having the intensities x and u presented simultaneously to the left and right eyes, respectively. Technically this is achieved using a stereoscope—middle panel in Figure 2—as detailed in the Apparatus section.

Ordering. The second primitive, $\succsim$, is the ordering of stimuli: the formalism $(x, u) \succsim (y, v)$ means that the stimulus (x, u) is judged to be at least as bright as (y, v). The indifference relation $\sim$ is defined by: $(x, u) \sim (y, v)$ if and only if both $(x, u) \succsim (y, v)$ and $(y, v) \succsim (x, u)$ hold. Importantly, the symbols $\succsim$ and $\sim$ are used rather than $\geq$ and $=$, because the latter refer to ordering of real numbers, whereas the former refer to psychological judgments; this means that $\succsim$ behaves similarly to the ordering $\geq$ of the real numbers. Moreover, it is assumed that the ordering agrees with physical intensity in the sense that if the intensity of the stimulus received by one eye is held constant, the binocular brightness varies monotonically with intensity increase/decrease in the other eye.

Monotonicity and Fechner's Paradox Monotonicity of ordering has been demonstrated empirically by, e.g., Levelt (1965). Figure 1 depicts the equi-brightness curve produced from data by a respondent who adjusted a luminance in one eye to match a standard when the experimenter varied the luminance in the other eye.

The monotonicity assumption holds for all but the conditions in which luminance in one eye is very small compared to that in the other eye. This condition creates the well-known Fechner's paradox: that when the luminance ratio of the left (right) to right (left) eye falls below a certain level the subjective sense of brightness increases (Fechner, 1861)—less energy giving rise to sensation of increased intensity. In Figure 1 this occurs when the luminance in one eye is 13% less than in the other eye—Levelt's (1965) stimuli were structurally similar to those used here.

Luce's (2002, 2004) theory does not extend to the stimulus condition in which Fechner's paradox occurs. However, this does not present an important limitation. The crucial observation is that in binocular vision, the left and right eye share 70-80% visual input and more when looking into the distance. For this reason, it is rare for a natural binocular stimulus to exhibit the properties required to give rise to Fechner's paradox. In fact, as far as we know, Fechner's paradox has only been demonstrated by artificially separating the luminance sources to the left and right eye. As the theory under evaluation is a description of behavior, its not extending to this extreme viewing condition does not present a substantial limitation.

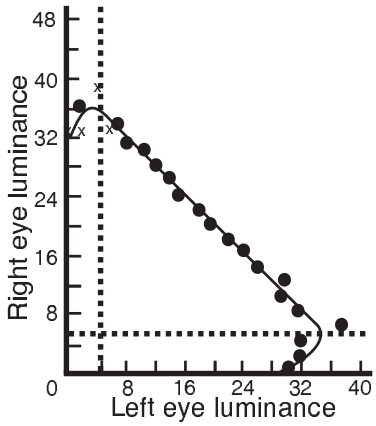

Figure 1. Reproduction of part of Fig. 5 of Levelt (1965). Shown is an equal-brightness curve for an observer. The experimenter adjusts the luminance in one eye and the respondent in the other to maintain a match to a standard.

Matching operation. A standard method in psychophysics (see Appendix A and, e.g., Stevens, 1975, for a comprehensive discussion) is that of matching a stimulus to a standard. We make use of this method and formalize it as follows. Each joint presentation (x, u) can be matched by (z, z), i.e., formally,

$$(x, u) \sim (z, z). \tag{1}$$

This is referred to theoretically as a *symmetric match.* Operationally, it will be referred to as *brightness matching* or just *matching* for short. It is convenient to express matching using an operator notion

$$x \oplus u := z, \tag{2}$$

where z is defined by 1 and the notation $A := B$ means A is defined by B. One can prove that $\oplus$ is, technically, a mathematical operator, which is referred to as the *summation operator.*

Note that (x, u) refers to a joint presentation of a signal pair. However, $x \oplus u$ refers to the cyclopic image that results. Since the cyclopic image results from the neural system's combining of the two input signals, we refer to this as a subjective *summation* of the two signals (nothing is hypothesized about how this summation is accomplished). In practice, the notations are used somewhat interchangeably.

The ordering primitive allows us to ask a natural question about brightness matching, namely about the *symmetry* of joint presentations:

$$x \oplus u \sim u \oplus x, \tag{3}$$

which is abbreviated as *jp-symmetry* (this is an analogue to the mathematical relationship: $a + b = b + a$). In words, is the image resulting from seeing stimulus of intensity x in the left eye and intensity u in the right eye subjectively judged the same as in the condition where the stimuli are switched?

This test amounts to asking whether the two eyes are behaviorally identical, and formally to asking whether the operator $\oplus$ is commutative. Whether or not this holds determines the next empirical steps, therefore this will be our first experiment.[1]

Generalized ratio production. A standard method in psychophysics (see Appendix A and, e.g., Stevens, 1975, for a comprehensive discussion) is that of magnitude or ratio production. Popularized by Stevens and used for decades, magnitude estimation entails the respondent producing a stimulus that is some proportion of a standard, e.g., two times as bright as a standard. The third primitive is a generalization of this magntitude/ratio production.

Suppose that $x > y \geq 0$ and let $p > 0$ be a positive number. Let (z, z) denote a signal pair that the respondent says makes the brightness "interval"[2] from (y, y) to (z, z) stand in the ratio p to the brightness interval from (y, y) to (x, x)—Figure 3. Clearly z is a function of x, y, and p. It is convenient to write this function as a mathematical operator of the following form: $(x, x) \circ_p (y, y) := (z, z)$. The *generalization* part of the ratio production can be seen by the fact that it agrees with ordinary *ratio production* when $(y, y) = (0, 0)$.

Notational convention. Let ϵ_l and ϵ_r denote thresholds for the left and the right eye respectively and let x' and u' be intensities of the stimuli presented in the left and the right eye, respectively; then our notation is $x = x' - \epsilon_l$ and $u = u' - \epsilon_r$. Thus, $x = 0$ denotes the threshold intensity (or less) of the left eye stimulus and $u = 0$ denotes the same for the right eye. For signals well above threshold, which are used, the difference $x - x'$ is negligible. Intensities are reported in cd/m^2.

Representations of $\oplus$ and $\circ_p$

Luce (2002, 2004) gave a set of necessary and sufficient (and testable) behavioral axioms, formulated in terms of the primitives, that allowed him to construct a numerical mapping Ψ, the psychophysical function, over the stimulus pairs that preserves the order $\succsim$, i.e.,

$$\Psi(x, u) \geq \Psi(y, v) \text{ iff } (x, u) \succsim (y, v), \tag{4}$$

and for which there exists a constant $\delta \geq 0$ such that

$$\Psi(x, u) = \Psi(x, 0) + \Psi(0, u) + \delta\Psi(x, 0)\Psi(0, u). \tag{5}$$

In addition, there is a strictly increasing numerical function W from the positive real numbers onto itself such that

$$W(p) = \frac{\Psi\left[(x, x) \circ_p (y, y)\right] - \Psi(y, y)}{\Psi(x, x) - \Psi(y, y)} \quad (x > y \geq 0). \tag{6}$$

[1]The original impetus for Luce's (2002) model of global psychophysical judgments was a series of results originally developed within the context of utility theory (Luce, 2000). They were based on an assumption of commutativity of this operator, which in the psychophysical context could be translated into saying that, e.g., the two eyes or ears are behaviorally identical. This hypothesis has been unambiguously rejected in audition (Steingrimsson & Luce, 2005a) and, as we shall see, for brightness as well (Experiment 1). The auditory data led Luce to generalize the result for the psychophysical context (Luce, 2002) and further in Luce (2004).

[2]The term "interval" is being used figuratively to refer to the difference in brightness that respondents experience between two intensity pairs.

In words, the order preserving condition (4) simply states that the brightness ordering judgment of the physical stimuli agrees with the physical intensity ordering. Property (5) captures the combining of inputs coming to the left and right eye, respectively, and is referred to as the *summation representation* (or sometimes as a *p-additive representation* in the literature). Property (6) describes the generalized ratio production operation and is referred to as the *production representation*.

There are two minor idealizations to note. First, these representations encompass the common idealization that x, u, and y are any non-negative real numbers. In practice, there is a maximum luminance level exposure that is safe for the human eye. Second, due to Fechner's paradox, the signal corresponding to, e.g., $(x, 0)$ is really equivalent to some (x', u') (evident from Fig. 1), but since the research domain is restricted to the stimulus space outside of where Fechner's paradox obtains, this is not of concern.

The important point is that the representations map a subjective input into a mathematical expression. That mathematical expression is indifferent to any biology and, in particular, nothing about any particular biology can be concluded from it. However, one strength is that it allows us to bring to bear the full force of the tools of mathematics to an unobservable subjective entity.

The representations consist of two unspecified functions Ψ and W, plus one constant. This allows for great freedom in capturing individual differences. Note that the functions and the representations are guaranteed to exist, provided that the parameter-free behavioral properties are satisfied. This situation is typical of axiomatic derivations of representations: no free parameters in the axioms and considerable freedom in the representations. Specifically, there is no fit of data to representations done, nor is any needed. For the reader who would like to delve into this issue, we recommend Appendix B as well as the comprehensive discussion of the general issue by R. Taagepera (2008) found in his descriptively titled book *Making Social Sciences More Scientific: The need for Predictive Models*.

Naturally, the actual mathematical forms of the unknown functions are of interest. In the axiomatic context, uncovering these forms entails formulating behavioral invariance properties that are equivalent to certain functional forms. Luce (2004) and Aczél and Luce (2007) have formulated such properties and Steingrimsson and Luce (2006, 2007) have evaluated them in loudness. For the psychophysical, Ψ, and the weighting function, W, respectively, they found support for power functions—or their generalization called Prelec functions—for most respondents. Future work aims at a parallel investigation for brightness. At this junction, we can only speculate about the forms of these functions. On the basis of decades of work on the form of the psychophysical function in a variety of domains (see Stevens, 1975, for an overview), that a power form might obtain would not be a surprise. Furthermore, given that W is a cognitive function (relating, in all domains, a number to an intensity), one might speculate that it is not domain specific, and hence, should a power or a Prelec form obtain, it would not be a surprise either. The form of the functions is of no concern for the present investigation.

Two behavioral properties of the representations

Now we turn to specifying the behavioral properties that will be tested and what can be concluded on the basis of their holding.

Subjective summation. From the results of Krantz, Luce, Suppes, and Tversky (1971, p. 250) the necessary condition of binary additive conjoint measurement, the *Thomsen condition*

$$\left.\begin{array}{l} x \oplus t \sim v \oplus w \\ v \oplus u \sim z \oplus t \end{array}\right\} \Longrightarrow x \oplus u \sim z \oplus w, \tag{7}$$

must hold. In a qualitative sense, the Thomsen condition describes the "additive cancellation" of t and z (easily seen by substituting $\oplus$ with $+$ and $\sim$ with $=$).

In the presence of some general background assumptions,[3] the Thomsen condition, (7), implies the summation representation, (5).[4] This property is evaluated in Experiment 2.

Production commutativity. The basic idea embodied in the representation of generalized ratio production, (6), is that the respondents perform the task as they are told to, i.e., to produce a brightness difference that is some multiple of the difference in brightness of a standard stimulus. An important and easily demonstrated (Luce, 2002, 2004) consequence of (6) is the behavioral property called *(subjective) production commutativity:* For $p > 0, q > 0$,

$$[(x, x) \circ_p (y, y)] \circ_q (y, y) \sim [(x, x) \circ_q (y, y)] \circ_p (y, y) \text{ where } y \geq 0 \tag{8}$$

Observe that the two sides differ only in the order of applying p and q, i.e., the test checks whether performing ratio production with p and then with q is equivalent to carrying out the same operation with the order of p and q switched. This is the reason for the term "commutativity" in its name. As previously mentioned, production commutativity with $y = 0$ also arose in Narens' (1996) theory. That hypothesis was sustained for brightness by Peißner (1999) using $p, q > 1$. In Experiment 3 we look at the case of $y > 0$.

The experimental program

The current experimental program is patterned on that carried out for loudness (Steingrimsson & Luce, 2005a,b; 2006; 2007). The first (current) paper tests the representations (5, 6) separately. This means that, a priori, the Ψ appearing in both representations is not guaranteed to be the same function. If we call them $\Psi_\oplus$ and $\Psi_{\circ_p}$, the second paper will deal with the question of whether $\Psi_\oplus = \Psi_{\circ_p} = \Psi$. The results of the current paper do not imply anything about this question. In a third paper, we will explore functional forms of the unknown functions, Ψ and W.

Experiments

Three experiments are presented: Test of jp-symmetry (3) (Exp. 1), the Thomsen condition (7) (Exp. 2), and production commutativity (8) (Exp. 3).

[3] Namely, monotonicity, solvability, and Archimedeanness (a way of stating that subjectively measured intensities are commensurable).

[4] The summation representation implies a stronger property called *double cancellation.* That property is the same as (7) were each $\sim$ replaced by $\succsim$. Obviously, double cancellation implies the Thomsen condition, but not the converse except when solvability and monotonicity obtain (see Krantz et al., 1971, for details). This is of no consequence here.

Common experimental methods

The experiments reported have a number of testing strategies in common that are now outlined. Other aspects are described later when relevant.

Respondents. A total of 21 students—graduate and undergraduate—from New York University and University of California, Irvine, and the author[5] participated in the three experiments in this article; although desirable, for practical reasons not all respondents participated in all of the experiments. All respondents reported normal or corrected-to-normal vision. All respondents, except the author, received compensation of $10 per session. Each person provided written consent and was treated in accordance with the "Ethical Principles of Psychologists and Code of Conduct" (American Psychological Association, 2002). Consent forms and procedures were approved by the Institutional Review Boards of New York University and UC Irvine.

Stimuli . The stimuli consisted of squares, subtending 10 degrees of visual angle, of achromatic light (RGB channels set to the same DAC value) displayed on a computer monitor located in a dark room—see the first panel in Figure 2 for an example.

Apparatus. Stimuli were generated with an Apple G4 using PsychToolbox extensions in MATLAB (Pelli, 1997; Brainard, 1997). At New York University, stimuli were presented on a 17" ViewSonic P810 CRT and at the University of California, Irvine, on an 18" NEC Multisync FE 950+ both with a resolution of 1024×768 pixels and refresh rate of 75 Hz. Experiments were conducted in a dark and light-insulated room.

Luminance calibration Equipment calibration and background conditions were of two kinds. A photometer, PhotoResearch PR-650, was used to measure luminance.

Condition 1 Calibration was done by averaging 5 repeated measures of luminance at every 5th of the 255 DAC values, starting at 1. Measures were taken from the left and right side of the monitor, equidistant from its center. The luminance measures were fitted to a Gamma function; the luminance disparity between sides was not appreciable.

Condition 2 The procedure in Condition 1 was changed to better deal with possible spatial inhomogeneity in the monitor luminance output.[6] A Gamma function was determined for each monitor location where a stimulus was displayed. One of these was picked as reference and DAC values for the other stimuli were determined to agree as closely as possible to a desired luminance using a reverse lookup procedure.

Background and luminance range/steps The monitors achieved an upper luminance of $\sim$100 cd/m^2 with the lowest stimulus level at $\sim$11 cd/m^2. Initially, stimuli were displayed on a no-luminance level (DAC value 0) background. In order to minimize the mixing of scotopic and photopic conditions, later experiments used 3.4 cd/m^2 background luminance,

[5]This we judged acceptable because knowledge of the experimental design has no influence on the behavioral tasks of matching and ratio production. The author is numbered as R8.

[6]Particular thanks go to M. Rudd who, as a reviewer, was responsible for this improvement in procedure.

a level at which photopic vision is dominant (R. Blake, personal communication, September 12, 2007). This change largely coincides with the use of calibration Condition 1. To maximize available adjustment options, all available DAC values were used.[7]

Stereoscope In two experiments, a stereoscope (see middle panel in Fig. 2) was used in the stimulus generation. A stereoscope uses a mirror system to project the left (right) half of the monitor to the left (right) eye. Placing a stimulus of intensity x (u) on the left (right) side, viewed through a stereoscope actualized the stimulus primitive (x, u)—see Section Joint presentations for details.

Procedure. Experiments were conducted in sessions lasting no more than one hour. The initial session was devoted to obtaining written consent, explaining the task, and running practice blocks. Depending on the experiment, practiced respondents typically completed around 60 estimates per session, organized into blocks of six or eight estimates, presented in randomized order. Rest periods were encouraged but their frequency and duration were under respondent control. Respondents received 10 minutes of dark adaption prior to each session. All respondents trained for one session on the task, except for Experiment 2 where training was three sessions—it was determined in pilot studies that respondents needed longer training before their data stopped showing large inter-session variability.[8]

The summation operation, $\oplus$, and matches The joint presentation (x, u), which can also be written as $x \oplus u$, means that the stimulus x is presented to the left eye and u to the right eye. The goal is to obtain estimates of the subjective brightness match of joint presentation, i.e., find the stimulus $z \oplus z$ that is perceived equal in brightness to $x \oplus u$, which can be written as $x \oplus u \sim z \oplus z$. Figure 2 describes the process: The first panel depicts what is displayed on the monitor, where the letters indicate stimulus intensity. The second panel depicts the stereoscope through which the respondents view the monitor (see section Stereoscopes in the Apparatus section). The third panel depicts what the subject sees. Since the stereoscope creates a cyclopic image, the percepts are those of $z \oplus z$ and $x \oplus u$.

To produce brightness matching, respondents adjust the intensity of z until they are satisfied that the two percepts—the upper and lower squares in the third panel of Figure 2—are equal in brightness. Specifically, respondents used key presses either to adjust the luminance of z or to indicate satisfaction with the brightness match. Respondents could chose any of four luminance steps 1, 2, 4, or 8 DAC values (extra-small, small, medium, large). After an adjustment, the screen was set to uniform background luminance for 100 ms. and then the next trial was presented—subjectively, this was experienced as a blinking and signaled that the adjustment had been made. This process was repeated until respondents were satisfied with the match, indicated by a key-press, at which time the trial

[7]Some researchers use linearized luminance steps. Since we seek subjective judgments from respondents and it is not a linear function of luminance, physical liniarization does not clearly provide an advantage over finer adjustments. Subjective linearization is problematic due to individual differences.

[8]The same phenomenon was observed in the analogous loudness experiment. The only hypothesis we have for the behavior is that this experiment is the only one in which intensity is adjusted in one ear/eye rather than both simultaneously, making the task-learning curve longer.

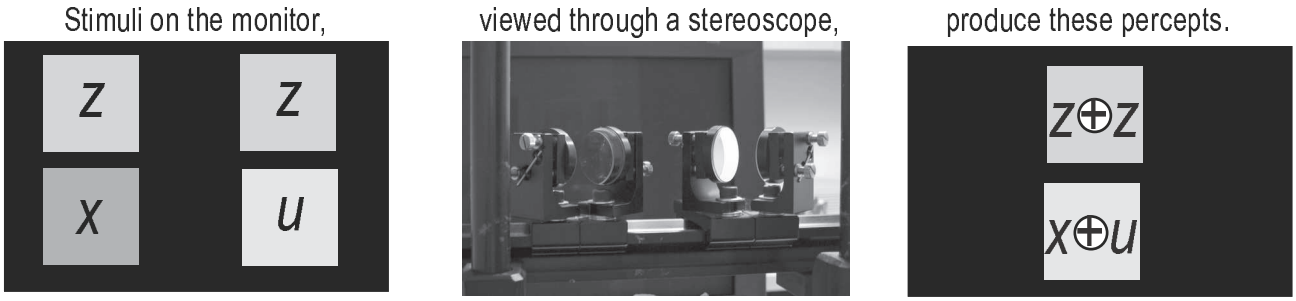

Figure 2. The first panel depicts the stimuli as displayed on the monitor. The second panel depicts the stereoscope through which respondents view the monitor. The third panel depicts the subjective percept seen by the respondents. The x, u, z are luminance values.

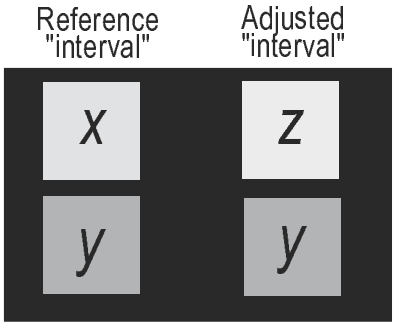

Figure 3. Stimulus in brightness ratio production. The x, y, z are luminance levels and p is a proportion. Respondents adjust the luminance of z until they are satisfied that the brightness interva between y and z is perceived as p times that between y and x.

ended and z was recorded as the response. Information about the current block and trial number were displayed in small letters in the upper left corner of the screen. In verbal instructions to respondents, the task was explained as that of making the upper stimulus equal in brightness to the lower one.

Ratio productions, $\circ_p$ The goal is to obtain the estimate $z \oplus z = (x \oplus x) \circ_p (y \oplus y)$ (in shorthand $z = x \circ_p y$). The task is to produce the intensity z such that the difference in brightness between y to z is a proportion p of the brightness difference between the reference signals y to x.

Figure 3 depicts the stimulus as displayed on the monitor and what the respondent looked at. The reference interval is on the left, from y to x. The adjusted interval is on the right, where y is reproduced and the respondent adjusts the intensity of z.

The adjustment procedure is the same as for matching; the proportion p was displayed on the upper left side of the monitor.

In instructions to respondents, the task was described as making the difference between the brightness of the lower and higher squares in the adjusted interval be, e.g., twice

($p = 2$) that between the reference squares on the left. Respondents were initially observed making the adjustments to help ensure complete understanding of the task.

Statistical method and presentation of results. Parameter-free null hypotheses of the form $L_{\text{side}} = R_{\text{side}}$, which reflect the behavioral properties, are tested. As a matter of logic, a null hypothesis can never be proven empirically, but it can be empirically supported. As is common in physics, the goal of the statistical test is to establish a criterion by which the data can be said to support (or not) the null hypothesis.

If the hypothesis $L_{\text{side}} = R_{\text{side}}$ is correct, it is equivalent to asserting that both L_{side} and R_{side} are drawn from the same distribution. Yet, because there is no theory that predicts the distributions of the estimates, a nonparametric test (Mann-Whitney U) is used for statistical evaluation, with a significance level of .05. This practice was used in similar studies (e.g., Ellermeier & Faulhammer, 2000; Zimmer, Luce, & Ellermeier, 2001; Ellermeier, Narens, & Dielmann, 2003; Zimmer, 2005; Steingrimsson & Luce 2005a, 2005b, 2006, 2007). Since intensity steps are discrete and estimates appear reasonably Gaussian, medians are best estimated by the mean, and variability by standard deviations. Hence these are the central tendency indicators reported. Because we do not have an a priori model of how individuals relate, all data analysis is done on individual data (e.g., Luce, 1995, p. 20).

To address the concern that sample sizes for L_{side} and R_{side} be sufficiently large to detect a true failure of the null hypothesis, all statistical results were verified using Monte Carlo simulations based on the bootstrap technique. The question asked is whether L_{side} and R_{side} could, at the .05 level, be argued to come from the same underlying distribution: 1000 resamplings of the data are made, each subjected to the Mann-Whitney U test, and the distribution of the results is examined—see Efron & Tibshirani (1993) and in particular Steingrimsson and Luce (2005a) for details. This is the criterion for accepting the null hypothesis as supporting the behavioral property.

Wilkinson and the Task Force on Statistical Inferences (1999) included the recommendation of adding "brief comments that place...effect sizes in a practical and theoretical context". Unfortunately, no practical guidelines were given for a non-parametric situation. Some effort has been made to tackle this problem but we are not aware of any satisfactory solutions at present time. One simple, yet powerful practical check on results lies in noting that should two medians (means) differ by less than Weber's fraction, they are arguably not noticeably different to an observer.[9] Teghtsoonian (1971) reports the mean Weber's fraction for brightness from five studies, all considered by him to be conservative, as 0.08. As the reader can verify, in only one case (noted in context and the result considered marginal) do we fail to reject a null hypothesis in a case where means differ by more than this fraction. In fact, in several cases, the null hypothesis is rejected when means differ by less than Weber's fraction, suggesting that the statistical criterion used for rejection is not lax.

Experiment 1: JP-Symmetry

Due to results in audition (Steingrimsson & Luce, 2005a), Luce (2004) expanded his theory (2002) to include the case where jp-symmetry fails. He termed this the "biased case"—the word "bias" simply means a deviation from jp-symmetry. Whether jp-symmetry

[9] We thank J. Yellott for this simple and elegant observation.

holds or not determines which of somewhat different sets of behavioral properties must be tested. Thus, this is the natural first property to test.

Method. Testing the property involves obtaining two types of matches: $x \oplus u = z \oplus z$ and $u \oplus x = z' \oplus z'$ and statistically testing whether $z = z'$. These matches are obtained as described in the section on the summation operation in the common methods section. Two luminance conditions were used, these are given in cd/m^2.

Luminance condition	a	b	c	Calibration
C_1	29.62	56.25	91.39	Condition 1
C_2	15.05	35.22	66.07	Condition 2

The three intensities, a, b, and c, gave rise to six ordered stimulus pairs: (a, b), (a, c), and (b, c) corresponding to the left side of (3), (b, a), (c, a), and (c, b) corresponding to the right side of (3), and three tests of the property. These six matching conditions were all run randomized within a block of trials.

Results. Fourteen individuals participated and their data are presented graphically in Figure 4. Each graph shows results of the six matching conditions, where, e.g., the matching of (x, u) is labeled xu (note: multiplication is not implied, this is a label only) and the rest analogously.

The statistical hypothesis is that $(x, u) \sim (u, x) \Leftrightarrow xu = ux$. Hence, the three statistical hypotheses to be tested are $ab = ba$, $ac = ca$, and $bc = cb$, which are marked on the abscissa. Average luminance level is marked on the ordinate. Sample size is indicated in the upper left portion of each graph.

The result of the statistical test is indicated on the abscissa, above the label of the relevant conditions, with no asterisk meaning the null hypothesis was not rejected at the 0.05 level, $\star$ denoting rejection at the 0.05 level, and $\star\star$ at the 0.01 level. Results from stimulus conditions C_2 are marked with as R#* above each graph.

Seven respondents (Rs 1, 3, 5, 6, 8, 29, 32) rejected all 3/3 testing conditions, two (Rs 27, 30) rejected 2/3 conditions, four (Rs 7, 9, 27, 30) rejected 1/3, and one (R2) rejected 0/3. Overall, the property was rejected in 29/42 tests.

For 13 of 14 respondents, one or more of the three conditions were found to be statistically different, with the trend for the remaining conditions consistent with this statistical difference. In 11/14 cases the pattern of results was that of $(x, u) \succ (u, x)$, which in terms developed by Luce (2004) is called a *right bias*. For two, weak evidence supports a *left bias* (Rs 34, 40). For the remaining respondent the hypothesis of no bias, i.e., $(x, u) \sim (u, x)$, was not rejected.

Luce's (2004) theory asserts nothing about the size or the direction of the deviation from symmetry (the bias). It does however assume monotonicity. Monotonicity would be violated if any of $ab < ac$, $ba < bc$, $ca < cb$, $ba < bc$, or $ca < cb$ should fail (R. Luce, personal communication, April, 2008). No violation of monotonicity is observed.

Discussion. The results echo those obtained in loudness: in general, jp-symmetry does not hold, but results differ in the sense that right bias is predominant (weak support

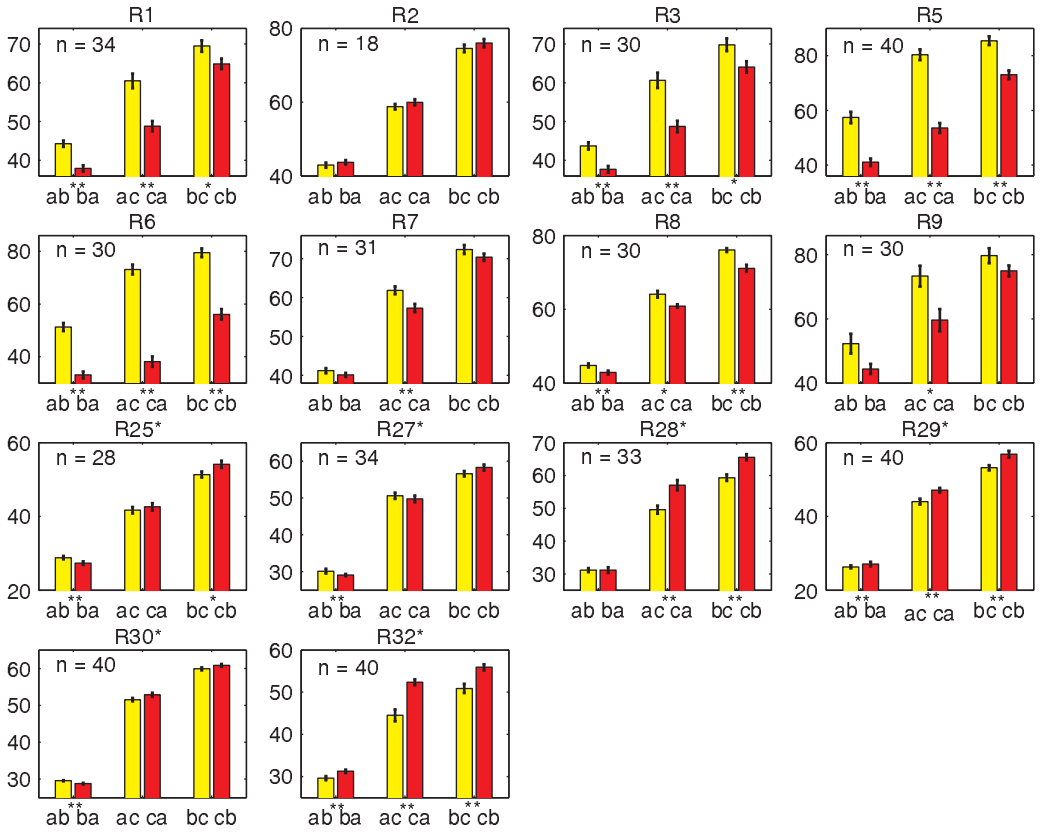

Figure 4. Experiment 1: Results from testing the jp-symmetry property, Eq. (3).

for two cases of left bias) where as a majority exhibited left bias in the auditory context (Steingrimsson & Luce, 2005a).

Luce's (2002, 2004) theory admits bias in either direction, but the theory makes no attempt to explain the proportions of people who are left or right biased. Dominance of side (left vs. right) is a common human feature. Beyond the familiar handedness, most people exhibit eye dominance; but whether such dominance plays a part in the bias observed is an open question. Steingrimsson and Luce (2005a) explored one explanation suggested to them, namely that differences in thresholds of the two ears could explain the bias behavior. They rejected that explanation based on the observation that in normal-hearing people, very small changes in energy were involved at the level of threshold compared to the energy in the well-above threshold stimuli used in their experiments. The disproportional energy differences are a consequence of loudness growing approximately as the power of intensity. Since the situation is analogous for brightness, the same argument applies here. Understanding the reasons for the failure of jp-symmetry is of interest but that is not the topic of this paper, nor is it necessary for any of our conclusions. Thus no further speculation is made with regard to this issue.

In conclusion, jp-symmetry (3) is not, as a general rule, found to hold for brightness.

Experiment 2: Thomsen condition

The goal of this experiment is to test the necessary condition of binary additive conjoint representation (Krantz et al., 1971, p. 250; Michell, 1990, pp. 68–73), the Thomsen condition, (7),

$$\left.\begin{array}{l} x \oplus t \sim v \oplus w \\ v \oplus u \sim z \oplus t \end{array}\right\} \Longrightarrow x \oplus u \sim z \oplus w. \tag{7}$$

To the author's knowledge, no test of the Thomsen condition (or related properties) have been carried out for brightness. Steingrimsson and Luce (2005a) studied several papers reporting support for the property in audition. They then devised a testing method that involved binaural stimuli and three estimation steps in which the intensity was adjusted in one ear only. This feature initially produced larger than expected variability in respondents' judgments, a feature that receded with additional training. Pilot studies with brightness replicated this phenomenon, and in response an additional four step process was devised in which judgments are more balanced in the sense that the trials feature conditions of adjusting luminance presented to the left eye only, right eye only, and both eyes. Results using both methods are reported.

Method. With reference to (7) and the notation of (1), the testing involved obtaining the estimates z', w', y' and y'' using two methods:

Method 1	Method 2
$(x,t) \sim (v,w')$	$(x,t) \sim (z',v)$
$(v,u) \sim (z',t)$	$(z',u) \sim (y',t)$
$(x,u) \sim (y',y')$	$(x,u) \sim (y'',v)$
$(z',w') \sim (y'',y'')$	

The testing requires that an estimate made in one trial be used as a stimulus in a subsequent trial. Steingrimsson and Luce (2005a) concluded that the best result was obtained when all estimates were collected within a session and the first estimate was used as the input in the subsequent estimation step. (See Appendices *A.1,2,* & *4* in Steingrimsson and Luce (2005a) for details.)

The Thomsen condition is said to hold if y' and y'' are not found to be statistically different.

All four/three trial types were run twice within a block in a pseudo-randomized order. Five stimulus conditions were used. These are given in cd/m^2 and were:

Stimulus Condition	x	t	v	u	Calibration	Method
C_1	44.58	25.31	11.48	69.29	Condition 1	1
C_2	43.11	24.43	10.88	66.91	Condition 1	1
C_3	34.21	29.91	17.28	58.36	Condition 2	2
C_4	26.09	23.09	13.28	40.49	Condition 2	2
C_5	25.95	35.22	40.49	30.37	Condition 2	2

Two observations for each estimate were collected within a block of randomized trials.

Results. Results for eight respondents are displayed in Table 1. In the table, averages, standard deviations, number of observations, and the results of the hypothesis tests are listed for each respondent. Luminance levels are in cd/m^2.

Resp.	Stimulus Condition	y' (s.d.)	y'' (s.d)	n	p_{stat}	Statistical Conclusion
R3	C_1	55.62 (3.66)	57.83 (8.07)	34	.189	$y' = y''$
R4	C_1	56.71 (2.30)	57.40 (4.31)	30	.917	$y' = y''$
R6	C_1	56.17 (2.15)	55.85 (2.63)	30	.557	$y' = y''$
R9	C_1	51.89 (2.76)	52.13 (5.45)	30	.976	$y' = y''$
R14	C_1	53.63 (4.20)	56.88 (9.49)	30	.242	$y' = y''$
R8	C_2	52.91 (2.75)	53.97 (4.53)	30	.445	$y' = y''$
R8	C_3	87.90 (5.00)	88.08 (5.40)	30	.841	$y' = y''$
R25	C_3	84.74(9.07)	79.07 (13.09)	30	.056	$y' = y''$
R29	C_3	70.81 (8.56)	73.14 (7.44)	30	.219	$y' = y''$
R8	C_4	65.96 (5.26)	63.78 (4.46)	30	.136	$y' = y''$
R29	C_4	50.27 (6.62)	50.98 (5.89)	30	.717	$y' = y''$
R25	C_5	49.77 (14.29)	44.35 (8.10)	30	.096	$y' = y''$

Table 1: Experiment 2: Results from testing the Thomsen condition, (7)

The property is not rejected for any of the 8 participants.

Discussion. This property is crucial to establish the summation representation, (5). Whether brightness summates has been discussed a fair bit and the conclusions have varied (see Appendix C). Additivity (as discussed in the Axiomatic literature, see Krantz et al., 1971 for details) implies that the brightness percept increases monotonically with luminance. Levelt (1965) tested this directly and showed just such monotonicity in the region where Fechner's paradox does not obtains (see Fig. 1). The current results suggest that this brightness increase is captured by the summation representation (5).

The reason for using two testing methods was the observation in pilot data that respondents required substantially more training than in other tasks in order for inter-session variability to stabilize—this is curiously parallel to loudness (Steingrimsson & Luce, 2005a). Alas we did not find method 1 to solve the problem and eventually reverted to method 2 which is quicker by virtue of having one fewer condition.

The property is found to hold for all 8 respondents, or in 12/12 conditions. However, the support is particularly weak for R25 with p values of .056 and .096, respectively and in the latter case, the means differ by slightly more than Weber's fraction of 0.08 (See the section on Statistics). We will follow the criterion set in the section on statistics, but the support is really only acceptable for 7/8 respondents and in 10/12 conditions. However, that is quite good, hence we conclude that the Thomsen condition has initial support in brightness.

Experiment 3: Production commutativity

Production commutativity, (8), is given by

$$[(x,x) \circ_p (y,y)] \circ_q (y,y) \sim [(x,x) \circ_q (y,y)] \circ_p (y,y), \text{ where } y < x.$$

In principle, one would want to test the property for a variety of values of p and q. In practice, the results of Steingrimsson and Luce (2007) suggest that the numerical distortion function, W, (6) differs for numbers above and below 1 and that, algebraically, nothing simple emerges from looking at the mixed case, e.g., where $p < 1$, $q > 1$. Consequently, in practice the case of $p < 1$, $q < 1$ and $p > 1$, $q > 1$ are considered separately.

Method. The testing requires four estimates in two steps. The first step consists of estimating v and then w in

$$(x,x) \circ_p (y,y) \sim (v,v),$$
$$(v,v) \circ_q (y,y) \sim (w,w),$$

and the second of estimating v' and then w' in

$$(x,x) \circ_q (y,y) \sim (v',v'),$$
$$(v',v') \circ_p (y,y) \sim (w',w').$$

Production commutativity is considered to hold if w and w' are found not to differ statistically. This experiment requires previous estimate to be used as input into a later estimate. Each individual estimate was used in subsequent estimates (see method section for Exp. 2 for more details). These matches are obtained as described in the section on ratio productions in the common methods section. Two observations for each estimate were collected within a block of randomized trials. The stimuli in cd/m^2 were:

Condition	y (UCI)	x (UCI)	p	q	Background
C_1: $p > 1, q > 1$	11.48 (10.88)	29.62 (28.62)	2	3	0 (3.4)
C_2: $p < 1, q < 1$	11.48 (10.88)	56.25 (54.37)	2/3	1/3	0 (3.4)

Results. Ten respondents participated in this experiment. One person participated in condition C_1 only, and two respondents (data not presented) sought repeatedly to adjust their responses beyond the upper luminance limit of the monitor, making the respondents' desired adjustments inaccessible. The remaining data are presented in Table 2.

The table lists respondent, condition, the means and standard deviations of w and w', and results of the statistical tests. Conditions marked * mean that data were collected using calibration Condition 2; other data used calibration Condition 1. Note that for R4 and R8, data were collected with both calibration methods. The only substantive difference is that while the property was rejected in C_1 using the initial calibration, it was not rejected the second time around.

Resp.	Cond.	w (s.d.)	w'(s.d.)	p_{stat}	n	Statistical Conclusion
R3	C_1	67.05 (18.74)	69.16 (15.00)	.643	32	$w = w'$
	C_2	23.36 (3.41)	22.28 (5.03)	.130	30	$w = w'$
R4	C_1	72.65 (12.56)	69.92 (14.28)	.226	32	$w = w'$
	C_2	17.82 (1.57)	19.38 (1.78)	$< .001$	30	$w \neq w'$
R8*	C_1	59.71 (5.84)	61.78 (5.83)	.189	28	$w = w'$
	C_2	17.74 (3.62)	17.03 (3.07)	.490	28	$w = w'$
R11	C_2	17.39 (6.27)	16.80 (3.73)	.979	32	$w = w'$
R20	C_2	19.80 (1.35)	20.27 (1.82)	.176	40	$w = w'$
R23	C_1	68.95 (6.49)	68.99 (7.24)	.733	30	$w = w'$
	C_2	29.17 (2.62)	29.55 (2.56)	.553	30	$w = w'$
R4*	C_1	71.39 (11.78)	69.28 (13.56)	.189	32	$w = w'$
	C_2	17.82 (1.57)	18.53 (1.78)	.305	30	$w = w'$
R8*	C_1	62.66 (6.44)	59.73 (5.37)	.221	32	$w = w'$
	C_2	17.74 (3.62)	17.03 (3.07)	.490	30	$w = w'$
R31*	C_2	70.40 (13,11)	71.65 (9.65)	.801	30	$w = w'$
	C_2	21.26 (3.23)	19.28 (5.33)	.230	30	$w = w'$
R33*	C_1	72.30 (9.03)	69.33 (9.91)	.121	30	$w = w'$
	C_2	17.97 (6.03)	16.80 (2.73)	.344	30	$w = w'$

Table 2: Experiment 3: Results from testing the proportion commutativity property, (8).

The property was accepted in 17/18 tasks.[10]

Luce (2002, 2004) assumes monotonicity. Monotonicity here would be violated should judgments of $p = 2 > q = 3$ and of $q = 1/3 > p = 2/3$. That prediction was examined and was not violated in a single case (data not presented).

Discussion. Peißner (1999) investigated the related property, threshold-production commutativity, namely,

$$[(x, x) \circ_p (0, 0)] \circ_q (0, 0) \sim [(x, x) \circ_q (0, 0)] \circ_p (0, 0),$$

which is the special case of (8) in which $y = 0$. He used an experimental paradigm and stimuli similar to those employed here, except he used only $p, q > 1$ and found it to hold. The property remains to be tested for $p, q < 1$. Here the generalized proportion commutativity is tested. Given that it is sustained in 17/18 cases, it is concluded that the Thomsen condition has acceptable initial support in brightness. This conclusion establishes the production representation, (6).

[10] R4's rejection for C_2 seems surprising: the individual passed the same property in audition and also in pilot studies for brightness (different equipment and evolving conditions). Yet, there was no objective reason to collect more experimental data so the result stands. The respondent did however fail to reject this condition for a a different set of stimuli.

Summary, conclusions, discussion, and further work

Summary and conclusions

The test results are summarized in Table 3.

Exp. #	Name	#R	#Tests	#Fail
1	JP-Symmetry	14	42	29
2	Thomsen condition	8	10(12)	0 (2 close)
3	Proportion. Commutativity.	8	18	1

Table 3: Summary of experimental results

The topic has been a theory of global psychophysical judgments leading to the two representation classes. With a failure of jp-symmetry (3), the asymmetric case, the theory leads to the two representations:

$$\Psi(x,u) = \Psi(x,0) + \Psi(0,u) + \delta\Psi(x,0)\Psi(0,u) \quad (\delta \geq 0), \tag{5}$$

$$W(p) = \frac{\Psi\left[(x,u) \circ_p (y,v)\right] - \Psi(y,v)}{\Psi(x,u) - \Psi(y,v)} \quad [(x,u) \succ (y,v) \gtrsim (0,0)]. \tag{6}$$

The aim of this article has been the separate testing of the properties derived from the first and second expression. Note that the above Eqs. (5, 6) are reproduced as presented by Luce (2004). However, the results of the present paper can, at best, support (5) and (6) with different psychophysical functions, $\Psi_{\oplus}$ and $\Psi_{\circ_p}$ (possibly but not necessarily the same). Although ongoing work, not reported here, strongly supports the hypothesis that $\Psi_{\oplus} = \Psi_{\circ_p}$, the overall conclusion from the three experiments reported is that the summation and production forms of Luce's (2002) theory are separately supported in the brightness domain. This conclusion is identical to that found by Steingrimsson and Luce (2005a) for loudness. Thus, Luce's (2002, 2004) theory seems to offer a certain unification on the level of description of at least these two domains. However, more work is needed before this statement can be made without reservation.

Further work

The overarching goal is to replicate the work of Steingrimsson and Luce (2005a,b; 2006; 2007) in brightness. The present paper parallels Steingrimsson and Luce (2005a) in form and result. This favorable outcome paves the way for a second paper to parallel Steingrimsson and Luce (2005b). That paper will evaluate the question of whether $\Psi_{\oplus} = \Psi_{\circ_p}$.

Another worthwhile research avenue is to extend the work to other types of stimuli, monitors with larger luminance ranges, chromatic stimuli, etc. Currently, all of these are being explored.

References

Aczél, J., & Luce, R. D. (2007). A behavioral condition for Prelec's weighting function on the positive line without assuming W(1) = 1. *Journal of Mathematical Psychology*,

51, 126-129.

American Psychological Association. (2002). Ethical principles of sychologists and code of conduct. *American Psychologist*, **57**, 1060–1073.

Anstis, S., & Ho, A. (1998). Nonlinear combination of luminance excursions during flicker, simultaneous contrast, afterimages and binocular fusion. *Vision Research*, **38**, 523–539.

Bolanowski, Jr., S. J. (1987). Contourless stimuli produce binocular brightness summation. *Vision Research*, **27**, 1943-1951.

Bourassa, S. J., & Rule, C. M. (1994). Binocular brightness: a suppression-summation trade off. *Canadian Journal of Experimental Psychology*, **48**, 418–434.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, **10**, 433-436.

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science*, **3**, 186-190.

Cohn, T. E., & Lasley, D. J. (1976) Binocular vision: Two possible central interactions between signals from two eyes. *Science*, **192**, 561–563.

Curtis, D. W., & Rule, S. J. (1978). Binocular processing of brightness information: A vector-sum model. *Journal of Experimental Psychology: Human Perception and Performance*, **4**, 132–143.

Ding, J., & Sperling, G. (2006) A gain-control theory of binocular combination. *Proceedings of the National Academy of Sciences*, **103**, 1141–1146.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* Chapman and Hall, New York.

Ellermeier, W., & Faulhammer, G. (2000). Empirical evaluation of axioms fundamental to Stevens's ratio-scaling approach: I. Loudness production. *Perception and Psychophysics*, **62**, 1505–1511.

Ellermeier, W., Narens, L., & Dielmann, B. (2003). Perceptual ratios, differences, and the underlying scale. In B. Berglund and E. Borg (Eds.) *Fechner Day 2003. Proceedings of the 19th annual meeting of the International Society for Psychophysics.* Stockholm, Sweden: International Society for Psychophysics. Pp. 71–76.

Engel, G. R. (1969). The autocorrelation function and binocular brightness mixing. *Vision Research*, **9**, 1111-1130.

Falmagne, J.-C. (1976). Random conjoint measurement and loudness summation. *Psychological Review*, **83**, 65–79.

Falmagne, J.-C., Iverson, G., & Marcovici, S. (1979). Binaural "loudness" summation: Probabilistic theory and data. *Psychological Review*, **86**, 25–43.

Fechner, G. T. (1861). Über einige Verhältnisse des binocularen Sehens. *Abhandlungen der mathematisch-physischen Classe der königlich sächsischen Gesellschaft der Wissenschaften, 5*, 337–564.

Gigerenzer, G., & Strube, G. (1983). Are there limits to binaural additivity of loudness? *Journal of Experimental Psychology: Human Perception and Performance*, **9**, 126–136.

Grossberg, S. (1997). Cortical dynamics of three-dimensional figureground perception of two-dimensional pictures. *Psychological Review*, **104**, 618–658.

Grossberg, S., & Kelly, F. (1999). Neural dynamics of binocular brightness perception. *Vision Research*, **39**, 3796–3816.

Irtel, H. (1998). Binocular brightness combination: A mechanism for combining two sources of rather similar information. In W. G. K. Backhaus, R. Kliegl & J. S. Werner (Eds.) *Color Vision: Perspectives from different disciplines.* Berlin: Walter de Gruyterp, 267–274.

Irtel, H. (1986). Experimente zu Fechners Paradoxon der binokularen helligkeit. *Zeitschrift für Experimentelle und Angewandte Psychologie*, **33**, 413–422.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement.* Vol. 1. Academic Press, New York.

Lehky, S. R. (1983). A model of binocular brightness and binaural loudness perception in humans with general applications to nonlinear summation of sensory inputs. *Biological Cybernetics*, **49**, 89–97.

Legge, G. E. (1984). Binocular contrast summation II. Quadratic summation. *Vision Research*, **24**, 385–394.

Levelt, W. J. M. (1965). Binocular brightness averaging and contour information. *British Journal of Psychology*, **56**, 1–13.

Levelt, W. J. M., Riemersma, J. B., & Bunt, A. A. (1972). Binaural additivity of loudness. *British Journal of Mathematical and Statistical Psychology*, **25**, 51–68.

Luce, R. D. (1995). Four tensions concerning mathematical modeling in psychology. *Annual Reviews of Psychology*, **46**, 1–26.

Luce, R. D. (2000). *Utility of gains and losses: Measurement theoretical and experimental approaches.* Erlbaum, Mahwah, N.J., errata: see Luce's web page at http://www.imbs.uci.edu/ personnel/luce.

Luce, R. D. (2002). A psychophysical theory of intensity proportions, joint presentations, and matches. *Psychological Review*, **109**, 520–532.

Luce, R. D. (2004). Symmetric and asymmmetric matching of joint presentations. *Psychological Review*, **111**, 446–454.

Luce, R. D., & Krumhansl, C. L. (1988). Measurement, scaling, and psychophysics. In: Atkinson, R. C., Herrnstein, R. J., Lindzey, G., Luce, R. D. (Eds.), *Stevens' handbook of experimental psychology*, 2nd Edition. Vol. 1 and 2. Wiley, New York, pp. 3–74.

Luce, R. D., & Narens, L. (1993). Comments on the "non-revolution" in the representational theory of measurement. *Psychological Science*, **4**, 127-130.

MacLeod, D. I. A. (1972). The Schrodinger equation in binocular brightness combination. *Perception, 1*, 321–324.

Michell, J. (1990). *An Introduction to the Logic of Measurement.* Erlbaum, Hillsdale, N. J.

Narens, L. (1985). *Abstract measurement theory.* Cambridge, MA: MIT.

Narens, L. (1996). A theory of ratio magnitude estimation. *Journal of Mathematical Psychology*, **40**, 109–129.

Narens, L. (2002). A meaningful justification for the representational theory of measurement. *Journal of Mathematical Psychology*, **46**, 746–768.

Narens, L. (2006). Symmetry, Direct Measurement, and Torgerson's Conjecture. *Journal of Mathematical Psychology, 50*, 290–301.

Peißner, M. (1999). Experimente zur direkten Skalierbarkeit von gesehenen Helligkeiten. *Unpublished master's thesis*, Universität Regensburg.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, **10**, 437–442.

Poletiek, F. (2001). *Hypothesis-testing behaviour.* Psychology Press.

Roberts, F. S. (1979). *Measurement theory.* Vol. 7 of Encyclopedia of Mathematics and Its Applications.Reading, MA: Addison-Wesley.

Schneider, B. (1988). The additivity of loudness across critical bands: A conjoint measurement approach. *Perception & Psychophysics*, **43**, 211–222.

Schrodinger, E. (1926). Die Gesichtsempfindungen. In *Mueller–Pouillets Lehrbuch der Physik, Book 2, Part 1* (11th ed., pp. 456–560). Vieweg: Braunschweig.

de Silva, H. R., & Bartley, S. H. (1930). Summation and subtraction of brightness in binocular perception. *British Journal of Psychology*, **20**, 242–252.

Steingrimsson, R., & Luce, R. D. (2005a). Evaluating a model of global psychophysical judgments: I. Behavioral properties of summations and productions. *Journal of Mathematical Psychology*, **49,** 290–307.

Steingrimsson, R., & Luce, R. D. (2005b). Evaluating a model of global psychophysical judgments: II. Behavioral properties linking summations and productions. *Journal of Mathematical Psychology*, **49**, 308–319.

Steingrimsson, R., & Luce, R. D. (2006). Empirical evaluation of a model of global psychophysical judgments: III. A form for the psychophysical function and intensity filtering. *Journal of Mathematical Psychology*, **50**, 15–29.

Steingrimsson, R., & Luce, R. D. (2007). Empirical evaluation of a model of global psychophysical judgments: IV. Forms for the weighting function. *Journal of Mathematical Psychology*, **51**, 29–44.

Stevens, J. C. (1967). Brightness function: Binocular versus monocular stimulation. *Perception & Psychophysics*, **2**, 451–454

Stevens, S. S. (1975). *Psychophysics: Introduction to its perceptual, neural, and social prospects*. Wiley, New York.

Taagepera, R. (2008). *Making social sciences more scientific: the need for predictive models*. Oxford University Press, New York.

Teghtsoonian, R. (1971) On the exponents in Stevens' law and the constant in Ekman's law. *Psychological Review*, **78**, 71–80

Wade, N. J., & Ono, H. (2005). From dichoptic to dichotic: historical contrasts between binocular vision and binaural hearing. *Perception & Psychophysics*, **34**, 645–668.

de Weert, C. M. M., & Levelt, W. J. M. (1974). Binocular brightness combinations: Additive and nonadditive aspects. *Perception & Psychophysics*, **15**, 551–562.

Wilkinson and the Task Force on Statistical Inferences (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, **54**, 594–604.

Zimmer, K. (2005). Examining the validity of numerical ratios in loudness fractionation. *Perception & Psychophysics*, **67**, 569–579

Zimmer, K., Luce, R. D., & Ellermeier, W. (2001). Testing a new theory of psychophysical scaling: Temporal loudness integration. Fechner Day 2001. *Proceedings of the 17th Annual Meeting of the International Society for Psychophysics*. Lengerirch, Germany: Pabst.

Appendix A
Historical context

Psychophysicists typically study the relationship between subjective attributes and the physical stimuli from which they arise, in particular how this changes as a function of changes in intensity. This is a measurement approach that is not only concerned with the assignment of numbers to sensations, but also with the formal properties of the number system into which the observed psychological measures are being mapped. If the enterprize is successful, general laws that systematically relate sensations to physical attributes emerge. This is, e.g., embodied in the famous laws of Fechner and Weber as well as in the power law of Stevens.

Central to Stevens' approach was his method of magnitude estimation/production where respondents were instructed to either give a number in response to a stimulus or to produce a stimulus that invoked a sensation that is in some prescribed proportion to a standard stimulus. The data collected using these methods suggest that for intensive continua (brightness, loudness, pain, heat, cold, etc.) sensation grows approximately as a power function of physical intensity (Stevens, 1975, for summary and references). These conclusions have almost exclusively involved averaging data over respondents, fitting the results to functional forms, and then evaluating the goodness of fit, where the power function form emerged as a well-fitting one.

Also studied in psychophysics is how intensity summates when, e.g., signals are administered independently to the two ears (Levelt, Riemersma, & Bunt, 1972; Falmagne, 1976; Falmagne, Iverson, & Marcovici, 1979; Gigerenzer & Strube, 1983; Schneider, 1988; Steingrimsson & Luce, 2005a,b) or the two eyes (De Silva & Bartley, 1930; Levelt, 1965; Stevens, 1967; Engel, 1969; de Weert & Levelt, 1974; Cohn & Lasley, 1976; Curtis & Rule, 1978; Bolanowski, 1987; Irtel, 1998; Bourassa & Rule, 1994; Grossberg & Kelly, 1999; Ding & Sperling, 2006), and how the two modalities compare (Lehky, 1983; Wade & Ono, 2005). Summation has been studied in a variety of ways, all of which can be said to have in common that some comparison is made between sensation magnitudes when a signal intensity is varied in the two sensory organs. An example is *sensation matching*, which is the special case of ratio production where the proportion is one. Respondents are typically instructed to produce a stimulus equal in subjective intensity to that of a standard. In the studies here the standard involves, e.g., presenting lights of different intensities to the left and the right eye respectively and the respondent then *matches* the resulting sensation by adjusting the intensity of two other lights, also presented to the left and right eye respectively.

Few would deny the tremendous contribution that Stevens made to psychophysics. However, his approach has not been spared criticism. For present purposes, it suffices to mention that Stevens never seemed to have articulated or tested fundamental assumptions inherent in his magnitude estimation/production tasks. He appeared aware of some oddities that defied explanation, e.g., he discussed what he termed a "regression effect," namely that magnitude estimations produced fit to power functions that had slightly different exponents than did magnitude productions. He appeared to have assumed that respondents gave responses that preserved ratios and treated numbers in a veridical fashion. Lastly, he largely ignored individual differences, taking them to be "noise" in what he must have assumed was a universal mechanism (Stevens, 1975).

While this and similar approaches by those who may be called *scalers* have produced enormously useful information in a rather simple fashion, the problems just highlighted, as well as others not mentioned, can be seen as motivation for the approach taken by the so-called *axiomatizers*, in whose form the theory being evaluated here is forged (Luce & Krumhansl, 1988). These methods are detailed in the following appendix.

Appendix B
Theoretical background

In defense of the approach

In his book,Taagepera (2008) discusses the prominent role statistics play in the social sciences. He observes that although the laws of physics were discovered without statistical hypothesis testing, such testing is a central tool in the social sciences. This observation led him to ask whether the laws of physics, such as the law of gravity ($F = GMm/r^2$), might be uncovered by the statistical methods typically used by social scientists. To pursue this question, he produced a synthetic data set that fit the law of gravity (with a small random error) and sent it to 38 social scientists asking them if they could make sense of the relationship of the output variable y to the input variables, x_1, x_2, x_3 (he indicated having an idea of the relationship but refrained from telling about it so as not to influence the outcome—see p. 19 for precise wording). Though eight individuals responded, none managed to uncover the law of gravity. However, all 8 did uncover high correlations of a variety of kinds, reporting R^2's from .7 to .9 and even, in one case, as high as .98 (Chapter 2). Taagepera notes that although these results are quite satisfactory by current social science norms—the positive and negative correlations were correctly identified, input factors were significant, and R^2 was high—all failed to uncover the underlying pattern. This failure is likely to be due to their never having considered that a non-linear relationship might underlie the data, in large part because the high R^2 did not seem to motivate a need for further analysis. Most laws in physics are non-linear, yet in psychology, tests assuming linearity (e.g, ANOVA) are often used to the exclusion of exploring other possibilities. While some reservations may be attached to Taagepera's "experiment", he raises the intriguing possibility that "[i]f some social phenomena did follow quantitative laws of the format most frequent in physics...then the quantitative methods currently dominant in social sciences just might not suffice to discover them" (p. 20).

Comparing the kind of hypothesis testing typical in physics to that in psychology, it becomes clear that in physics predictions of precise relationships are evaluated, entailing the non-rejection of a null hypothesis, whereas in contrast a typical experiment in psychology involves a hypothesis that some variable(s) is not a factor and then proceeds to reject that null hypothesis (Taagepera, 2008, and, e.g., Poletiek, 2001, Chapter 1, for comprehensive discussion).

There is an important difference between inferences that may be made on the basis of each of these two types of hypothesis testing. Support of a null hypothesis may assert exact relationships between variables and thereby provide for theories that are predictive and answer both how and how much a variable matters. In contrast, the information gleaned from rejection of a null hypothesis gives at best the information that a particular variable matters, but neither how, nor by how much, nor does it result in a predictive theory.

Without going into what has been a long and complicated discussion in the philosophy of science, which largely harkens back to philosopher Karl Popper's theory of falsification, let it suffice to say that the (powerful) dominance of the theory of falsification in the social sciences to the exclusion of all other approaches, including the one that has given us modern physics, is on the face of it going a bit far. Thus, we suggest that the use of predictive models should at least be considered in the toolbox of the social sciences, and

that the current theory being evaluated here represents one use of this tool.

The axiomatic measurement approach

Previous theoretical work in axiomatic psychophysics has used operations analogous to either the summation or production operations (e.g., Levelt et al., 1972; Narens, 1996). However, the conceptual novelty in the psychological context (although typical in physics) is the linking of these two operations together. This linking proved critical for establishing that a common psychophysical function can be used to represent both summation and production (currently being worked on in brightness).

The following is a much abbreviated account of the basic steps of axiomatic psychophysics (for a more extensive treatment see, e.g., Krantz et al., 1971; Narens, 1985; & Roberts, 1979). The axiomatizers tend to treat the following type of problem:

> If a body of (potential) qualitative observations satisfies certain primitive laws—axioms that capture properties of these observations—then is it possible to find a numerical structure that accurately summarizes these observations? In technical terms, the question is: To which numerical structures is the set of qualitative observations isomorphic? An isomorphism is a one-to-one mapping between structures under which the structure of the one maps into that of the other. It is also desirable to have an explicit process whereby the numerical structure can be constructed from the qualitative one. (Luce & Krumhansl, 1988, p. 5)

The axiomatic approach to this problem can be outlined as follows:

> [M]easurement theory proceeds in the deductive fashion of mathematics: certain formal properties are defined and theorems are proved. These theorems take the form of assertions that, if certain properties are true of the structure in question, then certain conclusions follow as a matter of pure logic. First, there are the primitive relations among attributes that determine both the measurement representations...Most results in measurement theory come in pairs. The first specifies conditions (axioms) under which it is possible to find numerical representation of the qualitative information. In other words, it formulates properties of a qualitative set of observations that are adequate for a certain kind of measurement system or scale to be appropriate. Such a result is called a representation theorem. The second type of result, called a uniqueness theorem, determines how unique the resulting measure or scale is. (Luce & Krumhansl, 1988, p. 7)

It should be noted that while this literature is purely mathematical, the choice of structures to be studied is much influenced by the intended application.

All measurement, whether within the axiomatic or the scaling tradition, begins with a method to study aspects of internal states such as brightness. One has at present few other options than to ask the respondents for an overt response, e.g., which of two signals produces more or less of the psychological attribute in question (e.g., which is louder,

brighter, etc.).[11] For a variety of reasons, repeated questions of this kind tend to result in variability in responses, or what we take to be errors of measurement that must be dealt with statistically.

It should be noted that the axiomatic approach has not escaped criticism. Perhaps the most serious of these is simply that, despite decades of work, it has produced little in terms of practical results (e.g., Cliff, 1992), although Luce & Narens (1993) counter this criticism, pointing to applications in utility theory and elsewhere. What seems clear is that in the field of psychophysics, practical applications of the axiomatic measurement approach have begun to emerge. The progress is both in theoretical papers that set forward directly testable behavioral axioms (e.g., Narens, 1996, 2002, 2006; Luce, 2002, 2004) as well as in empirical papers that report results of such tests (e.g., Peißner, 1999; Ellermeier & Faulhammer, 2000; Zimmer et al., 2001; Ellermeier, et al., 2003; Zimmer, 2005; Steingrimsson & Luce 2005a, 2005b, 2006, 2007; current paper).

Recent directly relevant developments in axiomatic measurement

Narens (1996) set out what he thought had to be the implicit assumptions behind Stevens' magnitude estimation/production methods. The first was that respondents treated numbers in a veridical fashion. That is, if W denotes a function that describes a respondent's interpretation of numbers p, then Narens showed that Stevens must have assumed $W(p) = p$. He formulated a behavioral axiom which was equivalent to a slightly weaker demand, namely the power form $W(p) = p^b$. This property has been unambiguously rejected in audition (Ellermeier & Faulhammer, 2000; Zimmer, 2005; Steingrimsson & Luce, 2007) and in brightness (Peißner, 1999).

A second result of Narens (1996) was a behavioral axiom equivalent to asserting that measurements done using ratio productions are on a subscale of a ratio scale. This property has been tested for both loudness and brightness and has been generally sustained (Peißner, 1999; Ellermeier & Faulhammer, 2000; Zimmer, 2005). Luce (2002) extended this result of Narens (1996) to what he called generalized ratio production and formulated the equivalent (subjective) production commutativity, (8), which was sustained in loudness by Steingrimsson and Luce (2005a). It is tested here in Experiment 3.

Appendix C
Relation to previous work

The task of comparing current results to the existing literature would be a rather arduous one were it not for the paper of Grossberg and Kelley (1999). There is no compelling reason to do more than summarize their main conclusions; the reader interested in more detail on the literature is advised to start with Grossberg and Kelly (1999).

Grossberg and Kelly (1999) evaluated and created a taxonomy of all the binocular summation models they found to date (see their Table 1) and evaluated them against existing data. In addition they present an updated version of the FACADE theory (Grossberg, 1997).

[11] Recent development in functional magnetic resonance imaging and similar techniques may at some future date make it practical to observe neural activity in such a direct fashion that it becomes a meaningful direct measure of an internal response; for now we must rely on overt responses.

The history of brightness summation has touched on topics ranging from whether it summates at all to how different stimuli produce different percepts. In a 1930 paper, De Silva and Bartley wrote:

> Numerous results gathered from observations of stereoscopic phenomena such as retinal rivalry support the view that alteration of conditions upon one retina always exerts some influence on the functioning of the other retina. Consequently it is rather surprising that Fechner, Sherrington, Abney and Watson, Dawson, and others have claimed to have proved by experiment that there is no binocular summation of brightness...as the weight of argument from the standpoint of numbers of investigations, prestige of investigators and use of refined apparatus is decidedly against it, binocular integration as regards brightness has come to be a generally discredited fact in the literature.
>
> It was the purpose of this investigation to attempt to put this problem of binocular summation of brightness to a critical test, and in particular to determine if possible why the experimental results should be so flatly contradictory. (pp. 242–243)

The author cannot help but find it somewhat ironic, some 77 years later, to be bringing up the question of brightness summation anew. Suffice it to say that the conclusion of De Silva and Barley, the aggregate conclusions of Grossberg and Kelly's (1999) review of additional literature, and the conclusion of the current paper all support that, outside the bounds of Fechner's paradox, brightness summates but that the effect is generally small.

Large summation effects have only been asserted for stimuli consisting of Ganzfelds (stimuli of uniform luminance that cover the entire visual field). Bolonowski (1987) reported, using magnitude estimation in flashed Ganzfeld conditions, that "complete binocular brightness summation occurs." Bourassa & Rule (1994) reported two experiments, where the first used Ganzfeld conditions and the second used smaller targets with very low spatial frequencies. They found Ganzfeld stimuli "produced a large amount of binocular brightness summation and very little Fechner's paradox" whereas their smaller low-frequency stimuli "produced greater Fechner's paradox than the Ganzfelds, but more binocular summation and less Fechner's paradox than what is usually reported for small targets with abrupt contours."

Grossberg and Kelly (1999) identified 13 models, including the one they proposed. Four of those they classified as "eye-weighting," three as "vector summation," and six as "neural networks." Grossberg and Kelly (1999, Section 5) meticulously dissected each of the proposed models and found a number of limitations which are applicable to the current discussion. They argued that the eye-weighting models of Levelt (1965) and Engel (1969) suffered from using weights that do not allow for binocular summation. They further argued that a model introduced by de Weert and Levelt (1974) made predictions that are at odds with data. Their main problem with Irtel's (1986) model was that it did not seem to extend to Ganzfelds. This last argument seems too weak to reject the model outright. However, from the point of view of the current results, Irtel's (1986)[12] model relies on an invariance

[12]For a reference in English, see Irtel (1998). However, note a typo where $\varepsilon(tx)$ and $m(tx)$ at threshold should equal $\varepsilon(x)$ and $m(x)$, respectively (H. Irtel, personal communication, August 10, 2006).

condition which effectively assumes the two eyes are identical, an assumption rejected here in Experiment 1. Whether that is a problem easily fixed is not something that is addressed here.

Grossberg and Kelly (1999) discussed several vector summation models (Shcrodinger, 1926; MacLeod, 1972; Curtis and Rule, 1978; Legge, 1984) and found flaws with them all, mostly related to their extendability to various stimulus conditions such as Fechner's paradox. The latter criticism could also be leveled at the model tested here. Conversely, it seems going too far to reject these models on the grounds that they have limitations, especially when it is untested whether those limitations may be overcome, and stimulus conditions under which Fechner's paradox obtains are not likely in natural conditions. In contrast, from the point of view of the current work, it seems that none of the models explicitly treat the biased case, i.e., that the two eyes are different. It is not clear how hard it would be to amend these models to admit this empirical result, but for now this appears to be a flaw.

The rest of the models identified and discussed by Grossberg and Kelly (1999) are neural network models. While Grossberg and Kelly (1999) also found limitations in all the neural network models they evaluated, the main issue is, again, that these models treat the eyes as unbiased. However, it is not the intent here to reject every model proposed that reasonably accounts for data on the basis that they do not admit the biased case, especially as this situation may well be easily remedied.

Ding and Sperling (2006) is a recent entrant into the group of neurally inspired models. Their approach is to divide various biological processes into black boxes and to model their input-to-output functions. An explicit goal of this form of modeling is to make it possible to reach into any black box and break it up into smaller boxes and thereby, incrementally, fashion a model that continues to become more detailed and accurate. This modeling effort shares the goal of the current one of enabling the accumulation of results. The model appears to assume the eyes are identical but it is likely an assumption that is easily adjusted. Of course, as the model is inspired by specific biology, it is not directly extendable to other domains, but the methodology may become an import avenue to incrementally evolve biologically inspired models.

However, as emphasized in the second section of Appendix B, the goal is not to compete with or supplant existing models; the contrast drawn here is in the modeling approach. Luce (2002, 2004) takes an approach that in Grossberg and Kelly's (1999) and Ding and Sperling (2006) framework would define a new category, namely, an axiomatic model. Steingrimsson and Luce (2005a) articulated the main conceptual difference as follows:

> First, [Luce's (2002, 2004)]...theory is not domain specific in the sense that it can, in principle, apply to any intensive dimensions (e.g., loudness, brightness, or [heaviness]). Second, although neuronal activity ultimately underlies perception, the approach taken here is entirely behavioral. The important abstraction is that the results we obtain are valid answers to our questions regardless of what the neural machinery may be, only the behavior matters. In effect, we could use the very same approach to an alien life form or a robot. Consequently, we do not make any attempt to draw conclusions about the biological workings of the perceptual system from our results. (p.291)

For this reason, comparing neural network models to the model discussed here is not entirely apt.